Permutations Exercises

Stephen Blatti

August 7, 2017

R Markdown

```
library(downloader)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
url <- "https://raw.githubusercontent.com/genomicsclass/dagdata/master/inst/extdata/babies.txt"
filename <- basename(url)
download(url, destfile=filename)
babies <- read.table("babies.txt", header=TRUE)
bwt.nonsmoke <- filter(babies, smoke==0) %>% select(bwt) %>% unlist
bwt.smoke <- filter(babies, smoke==1) %>% select(bwt) %>% unlist
```

1. We will generate the following random variable based on a sample size of 10 and observe the following difference:

```
N=10
set.seed(1)
nonsmokers <- sample(bwt.nonsmoke , N)
smokers <- sample(bwt.smoke , N)
obs <- mean(smokers) - mean(nonsmokers)
```

The question is whether this observed difference is statistically significant. We do not want to rely on the assumptions needed for the normal or t-distribution approximations to hold, so instead we will use permutations. We will reshuffle the data and recompute the mean. We can create one permuted sample with the following code:

```
dat <- c(smokers,nonsmokers)
shuffle <- sample( dat )
smokersstar <- shuffle[1:N]
nonsmokersstar <- shuffle[(N+1):(2*N)]
mean(smokersstar)-mean(nonsmokersstar)
```

```
## [1] -8.5
```

The last value is one observation from the null distribution we will construct. Set the seed at 1, and then repeat the permutation 1,000 times to create a null distribution. What is the permutation derived p-value for our observation?

```
set.seed(1)
obsdiff <- mean(smokers) - mean(nonsmokers)
avgdiff <- replicate(1000, {
  all <- sample(dat)
  smokersstar <- all[1:N]
  nonsmokersstar <- all[(N+1):(2*N)]
  return(mean(smokersstar)-mean(nonsmokersstar))
})
pvalNull <- (sum( abs( avgdiff ) > abs( obsdiff)) + 1) / (length(avgdiff) + 1)
pvalNull
```

```
## [1] 0.05294705
```

```
#sln
# set.seed(1)
# null <- replicate(1000, {
#   shuffle <- sample( dat )
#   smokersstar <- shuffle[1:N]
#   nonsmokersstar <- shuffle[(N+1):(2*N)]
#   mean(smokersstar)-mean(nonsmokersstar)
# })
# ( sum( abs(null) >= abs(obs)) +1 ) / ( length(null)+1 )
# ##we add the 1s to avoid p-values=0 but we also accept:
# ( sum( abs(null) >= abs(obs)) ) / ( length(null) )
```

2. Repeat the above exercise, but instead of the differences in mean, consider the differences in median obs <- median(smokers) - median(nonsmokers). What is the permutation based p-value?

```
set.seed(1)
obsdiff <- median(smokers) - median(nonsmokers)
meddiff <- replicate(1000, {
  all <- sample(dat)
  smokersstar <- all[1:N]
  nonsmokersstar <- all[(N+1):(2*N)]
  return(median(smokersstar)-median(nonsmokersstar))
})
pvalNull <- (sum( abs( meddiff ) > abs( obsdiff)) + 1) / (length(meddiff) + 1)
pvalNull
```

```
## [1] 0.01798202
```